ASSESSING THE ACCURACY OF
NORMAL APPROXIMATIONS

by

JAMES S. HODGES
School of Statistics
University of Minnesota
Technical Report No. 449
June 1985

ABSTRACT
The widespread applicability and use of normal approximations creates a need for methods for assessing their accuracy in an operational fashion. The operational usefulness of two well-known existing approaches is discussed and two new methods are proposed and exemplified. These new methods are based on a comparison of the level curves of the exact function (density of likelihood) to the level curves of the normal approximation, and on a comparison of selected line integrals of the exact function and the normal approximation.

Key words: Approximations, Confidence regions, Curvature methods, likelihood regions, posterior density, predictive density

1. INTRODUCTION

Approximations based on the normal distribution are ubiquitous in statistics. Practitioners in all foundational schools use them: Bayesians use normal approximations for posterior densities (Zellner, 1971, pg. 32); the likelihood can often be approximated by a function proportional to a normal density; in large samples the distribution of the maximum likelihood estimate is approximately normal (Cox and Hinkley, 1974, pg. 294); and level curves of the likelihood (i.e. likelihood regions), which are approximate confidence regions, can in turn be approximated by ellipsoids (see section 4.2), which is equivalent to using a normal approximation to the likelihood.

With this widespread use comes a need to assess how well these approximations work, in some operational way. One such operational approach is through Fisher's disjunction (1973, section III-1). The disjunction is as follows: if an unexpected result is observed, then either something improbable has happened, or the postulated probability mechanism is not correct. This can be restated with a more data analytic flavor as: if some apparent feature of the data strongly contradicts your expectation, then either something improbable has happened, or one or more of the assumptions that led you to look at that feature of the data with a particular expectation is incorrect. Included in these latter assumptions is the assumption that any approximations used were accurate enough in the instance in question not to have caused the observed but unexpected result. To be useful, then, a method for assessing the accuracy of an approximation should answer the question of whether or not the approximate entity one has calculated displays a particular feature simply because the approximation has failed.

In section 2, two well known approaches to assessing the accuracy of approximations are considered from this point of view, and in the process three desirable qualities of such assessment methods are identified. In section 3, two new methods displaying these qualitites are developed, and in section 4 they are applied to approximate likelihood regions and to an approximate predictive density. Section 5 concludes with a comparison of these new methods to the well known methods of section 2.

2. ASSESSING THE ACCURACY OF APPROXIMATIONS

2.1 Rates of Convergence

The prevailing mode of thinking about the quality of asymptotic approximations is centered on characterizing the rate at which the error in the approximation diminishes as the size of one's sample increases. A typical statement in this mode is from Durbin (1980), pg. 311: "Our basic approximation for the density of T_n at $T_n = t$ [a function of the observable y] is

$$(\frac{n}{2\pi})^{m/2} |I(t)|^{1/2} \frac{f(y,\theta)}{f(y,t)} \{1 + O(n^{-1})\}," \qquad (2.1)$$

where the statement $g(y,\theta) = \hat{g}(y,\theta)\{1 + O(n^{-a})\}$ means that there exists a constant k (almost always unknown and depending on g, y, and θ) and a number N(k) (apply previous comment) such that

$$\frac{g(y,\theta)}{\hat{g}(y,\theta)} - 1 \leq \frac{k}{n} \qquad (2.2)$$

for all $n > N(k)$ (see, e.g., Bishop, Fienberg, and Holland, 1975, pg. 458).

2

In this mode of thinking, work centers on finding approximations with errors $o_p(n^{-a})$, $O_p(n^{-a})$, or their nonstochastic counterparts, for a as large as possible, because then these approximations will converge to the exact function or statistic most rapidly. Considerable effort, ingenuity, and serendipity have gone into methods for increasing a; see, for example, Welch and Peers (1963), who find an $o_p(n^{-1})$ method as an improvement on an $o_p(n^{-1/2})$ method, or Bartlett (1953a, 1953b), who gets the same order improvement for the same problem a different way.

Certainly it is very useful to think about asymptotic approximations this way. A method with $o_p(n^{-2})$ errors should be scrutinized for general use before a method with $o_p(n^{-1})$ errors that is equally difficult to use. But for a given model, a given data set, and a given approximation, this prevailing mode of thinking does not provide relevant information about the size of the errors one risks in using the approximation to make the usual sort of statistical statement. That is, while it is true that Durbin's approximation satisfies (2.2) for suitable models, as long as k and N(k) are unknown -- and k could conceivably be 10^{10} or 10^{-10} for the number of observations you have -- this is cold comfort indeed.

An analogy can be drawn to the Bayesian criticism of pre-test criteria for evaluating estimators or hypothesis tests. The essence of that criticism is that although it is true that a procedure that is good according to some frequentist criterion will have desirable properties "pre-test," i.e. averaged over some hypothetical sequence of repetitions, it may also be true that for any or all particular realizations of the assumed random process the procedure may give an undesirable or even plainly ridiculous result (see Lindgren, 1976, pp.

266-267 for an example). In judging asymptotic approximations, it is likewise true that while an approximation having $o_p(n^{-1})$ errors is promising, in any particular instance in which one might wish to use it the errors could be quite large for the sample size which will be available.

As an illustration of the nature of the problem, consider the single parameter inverse Gaussian regression model, where the observable $y > 0$ has density

$$(\frac{\lambda}{2\pi})^{1/2} \, y^{-3/2} \, \exp(\lambda y[x\beta] + \lambda(-2[x\beta])^{1/2} - \frac{\lambda}{2y}), \qquad (2.3)$$

where $\lambda > 0$ is a known parameter, $x < 0$ is a known observable regressor, and $\beta > 0$ is an unknown and unobservable regression parameter (see Johnson and Kotz, 1970, pg. 139; McCullagh and Nelder, 1983, pg. 22). Note that this is a flat exponential family. For this model, the m.l.e. for β based on n observations is

$$\hat{\beta} = \frac{x_0^2}{2(\sum\limits_{i=1}^{n} x_i y_i)^2}, \text{ for } x_0 = \sum\limits_{i=1}^{n} (-x_i)^{1/2}.$$

The exact distribution of $\hat{\beta}$ cannot be obtained explicitly, but for a given true value of β it can be obtained numerically by inverting the known characteristic function of $\sum x_i y_i$ and transforming the resulting density in the obvious way. This distribution depends only on x_0 and λ, and not on n explicitly. The usual normal approximation to the distribution of $\hat{\beta}$, assuming β_0 to be the true value, is normal with mean β_0 and variance $(4/\lambda x_0)b^{3/2}$, where b is the observed value of $\hat{\beta}$. As for the exact distribution, this approximate distribution does not

depend on n explicitly, but it does depend on the value of $\hat{\beta}$ which happens to have been observed, b. This latter effect can be quite large, as demonstrated by Figure 2.1. In Figure 2.1, the solid line is the exact density for $\hat{\beta}$, and the broken lines are the usual normal approximations, for b = 1.0, 5.0, and 9.0 (in order of decreasing peakedness), for $\lambda = x_0 = 1$ and true value $\beta_0 = 5.0$. The exact cumulative distribution function for $\hat{\beta}$ evaluated at 1.0, 5.0, and 9.0 are 0.055, 0.38, and 0.56, respectively -- not extreme values, and yet these approximations have variances of 4.0, 44.72, and 108.00 respectively, and give 1.72 and two negative numbers as approximate lower 0.05 quantiles, for which the exact value is 0.94.

In fact, since the exact and approximate distributions for $\hat{\beta}$ depend only on x_0, λ, and b, the following distressing fact follows: for the given true value $\beta_0 = 5$ and arbitrary given n, a pattern of x_i yielding the x_0 of this example can be found, so that the problem displayed in Figure 2.1 can occur for arbitrarily large n. This is true even if the x_i contain repetitions of some design--the problem is not a failure to converge, but slow convergence.

The two undesirable properties of convergence rates illustrated by this example--that they will not alert a user to an approximation that can vary substantially with the observed value of $\hat{\beta}$, and that they might not alert a user to a failure to converge for an arbitrarily large sample--arise because little or no information about the data and model at hand could be taken into account in deriving the convergence rate. The single most important property that accuracy assessment methods need to meet the task set in section 1 is that they must use this information.

Fig. 2.1--Exact and approximate densities for the inverse
Gaussian regression parameter estimator. Solid
line = exact density, dotted lines = approximate
densities
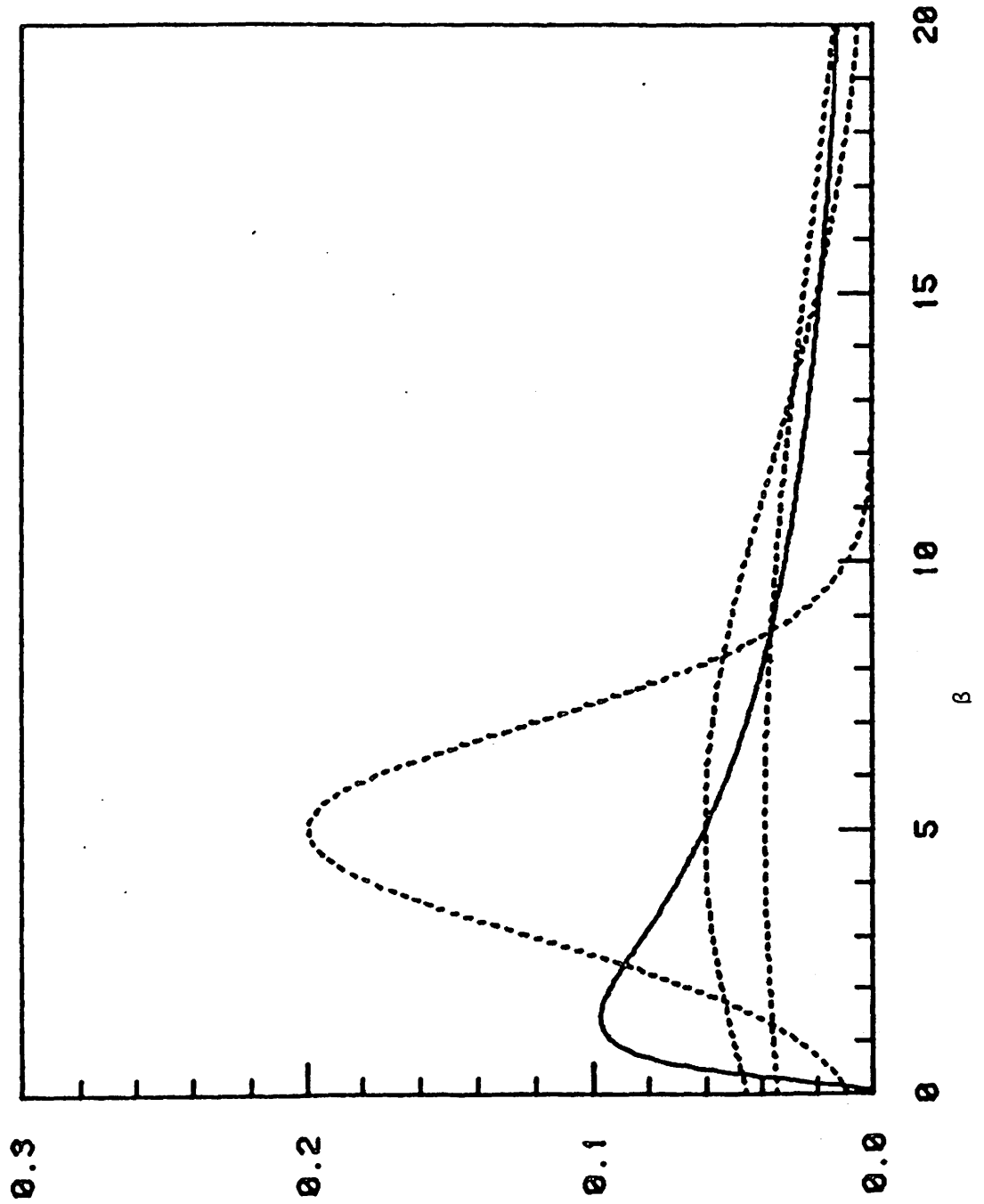
2.2 A geometric approach

In recent years, ideas from differential geometry have been applied to some types of models and asymptotic approximations in an attempt to use the information which the predominant method does not use. Beginning with Beale (1960), Efron (1975), and Bates and Watts (1980), differential geometric ideas such as curvature have been suggested as bases for "post-test" evaluations of the adequacy of asymptotic approximations, and for other purposes as well.

The approach of Bates and Watts (1980) has been most fully developed for practical use. This method is applicable to nonlinear regression models, i.e. models in which the observable y is postulated to be normally distributed with mean $F(x;\theta)$ and variance σ^2 independently of other such observable y's, where F is a known function of arguments x and θ, x is a known q-vector of observable regressors, θ is an unknown, unobservable p-vector of parameters, and $\sigma^2 > 0$ is unknown. For the usual normal linear model, q = p,

$$F(x;\theta) = \sum_{i=1}^{p} x_i \theta_i ,$$

and the exact distribution theory for this model is well known. For functions $F(x;\theta)$ that are nonlinear in θ, application of inferential methods is much more difficult, and tractable exact frequentist results are available only for special cases (e.g. Gallant, 1975). Because the added difficulty coincides with the nonlinearity of $F(x;\theta)$ as a function of θ, Bates and Watts' method depends on examination of curvature properties of the solution locus $\eta'(\theta) = (F(x_1;\theta),....,F(x_n;\theta))$. Bates and Watts (1980, 1981) and Hamilton, Bates, and Watts (1982) concentrate on local curvature properties around the m.l.e., which is also the least squares estimator for this model.

In the sequel, we assume that the reader is familiar with Bates and Watts' approach (see section 2 of Bates and Watts, 1980, or Ratkowsky, 1983), particularly the ideas of intrinsic curvature and parameter effects curvature, and the calibrations of the two curvatures.

This geometric approach, while a step in the right direction, has several problems as a method of assessing the accuracy of approximations. First, it does not appear to work as advertised. Cook and Witmer (1985) present examples in which (i) the parameter effects curvature is larger than the guide value, but the approximate likelihood region is very similar to the exact likelihood region, and (ii) the intrinsic curvature is identically zero, and some values of the m.l.e. produce large maximum parameter effects curvatures but eminently acceptable approximate regions, while other values for the m.l.e. produce apparently acceptable curvatures but approximate likelihood regions bearing little resemblance to the exact regions they are supposed to approximate. However, simulations done for the examples mentioned for case (ii) (Weiss, 1985) show that for this case, the actual coverage rate of the approximate likelihood region does depart substantially from the nominal rate when the maximum parameter effects curvature is larger than Bates and Watts' guide value, and that the actual coverage rate is related to maximum parameter effects curvature roughly by a smooth curve--even though Bates and Watts' development has no obvious connection to coverage rates.

This interesting result does point to two other difficulties with this geometric approach. The likelihood is determined by the solution locus, so derivatives of the solution locus, and thus the curvatures, convey information about the likelihood. But the derivatives of the solution locus are complicated

functions of the derivatives of the likelihood or log likelihood, and if we are interested in the latter, the information conveyed by the former may not be in a desirable form. In particular, it is difficult to think about how the curvatures should be calibrated; in practice, Bates and Watts' guide values seem to be useful mainly as rough guides.

Finally, the curvature methods are themselves defined by making an approximation, namely a quadratic Taylor series approximation to the solution locus. While this does not mean that the curvatures are necessarily uninformative or misleading, it does mean that to overcome the shortcomings of the Taylor series approximation will require a redefinition of the method and not simply a refinement.

2.3 Desirable properties of accuracy assessment methods

The examples in sections 2.1 and 2.2 illustrate properties that methods for assessing the accuracy of approximations should possess to be operationally useful:

(i) they should make efficient use of the available information about the function, region, or statistic being approximated,

(ii) they should be calibrated in a scale that permits them to be used, for example, in the revised Fisher disjunction of section 1, and

(iii) as much as possible, approximations should be avoided in defining and computing the methods.

3. TWO METHODS FOR ASSESSING THE ACCURACY OF NORMAL APPROXIMATIONS

Let f and g be two unimodal probability densities on a connected subset Ω of real p-space R^p, that are nonzero on the interior of Ω. If infinitely powerful costless computing were available, an obvious way to compare f and g would be to specify some subsets of Ω and compute the probabilities that f and g assign to those sets. Infinitely powerful costless computing is not available, and for our problem the density g will be assumed to be difficult to use in computations, so the comparison strategies must be different.

If f and g are "very similar," two things will be true: (1) the level curves or contours of f and g, $C(f,k) = \{x|f(x) = k\}$ and $C(g,k) = \{x|g(x) = k\}$ will be "very similar," and (2) the integrals (including line integrals) with respect to the two densities of functions $\ell(\cdot)$,

$$\int \ell(x) f(x) dx \quad \text{and} \quad \int \ell(x) g(x) dx,$$

will be "very similar." The two comparison strategies presented here are based on comparing these two features of the densities.

3.1 Contours

If $f(x) = g(x)$ for all x, then for a given k, the level curve $C(f,k)$ will closely resemble $C(g,k)$. More precisely,

$$d_1(k) = \max_{y \in C(g,k)} f(y) \quad \text{and} \quad d_2(k) = \min_{y \in C(g,k)} f(y)$$

will both be close to k; and if either d_1 or d_2 is far from k, f and g differ.

To illustrate, let p = 1 and consider the densities f and g in Figure 3.1. For these densities, $C(f,k) = \{x_1,x_2\}$, $C(g,k) = \{y_1,y_2\}$, and $d_1(k)$ and $d_2(k)$ are as shown.

How large is "large?" i.e. how are d_1 and d_2 to be calibrated? Suppose now that p = 1 and that in addition to the properties of f assumed at the beginning of this section, f(x) is the known symmetric density of a random variable X having μ as its mode. Then for any k, C(f,k) consists of two points equidistant from μ, and the probability assigned by f to the interval between these two points can be calculated (for a density such as the normal density, this is very easy). Thus, to each k there corresponds a probability computed using f,

$$p_0(k) \equiv \int_{x_1}^{x_u} f(u)\ du = Pr_f(f(X) \geq k),$$

where the numbers $x_u \geq x_l$ are the points in C(f,k) and the subscript on "Pr" indicates the density used to compute the probability. The numbers d_1 and d_2 can be calibrated using this idea. Compute the two probabilities

$$p_1(k) \equiv Pr_f(f(X) \geq d_1(k))\ \text{and}\ p_2(k) \equiv Pr_f(f(X) \geq d_2(k)).$$

If these probabilities differ greatly from $p_0(k)$—where "greatly" depends on the user's intent—then the densities f and g may be said to differ, for the user's purposes.

To extend the method to p-dimensional random variables, for p > 1, consider lines in p-space of the form $\mu + th$, for h a unit length direction vector, μ the mode of f, and t a real number. Again, assume that f is the density of a p-

10

dimensional real random variable X with the properties listed at the beginning of the section. To generalize the requirement that f be symmetric, we require that $f(\mu + th) = f(\mu - th)$ for $t > 0$, for every h (this includes the multivariate normal and Student t distributions). Then for $k < \max f(x)$, the two members of $C(f,k)$ in the direction h from μ--call the pair of members $C_h(f,k)$--will correspond to two scalar multipliers $s_1 = -s < 0$ and $s_2 = s > 0$, so that $f(\mu \pm sh) = k$. Similarly, the members of $C(g,k)$ in the direction h from μ--call them $C_h(g,k)$--will correspond to scalar multipliers t_j so that $g(\mu \pm t_j h) = k$. Note that because g is unimodal, $C_h(g,k)$ can have zero, one, or two elements, depending on the location of g's mode relative to f's mode, even if $k < \max g(x)$. Given a direction h and f's mode μ, f and g may be compared along h as in the single dimensional case. Define

$$d_1(k,h) \equiv \max_{y \in C_h(g,k)} f(y) \quad \text{and} \quad d_2(k,h) \equiv \min_{y \in C_h(g,k)} f(y);$$

if d_1 or d_2 differs much from k, f and g differ in the direction h. The calibration of d_1 and d_2 is extended to the multidimensional case by defining

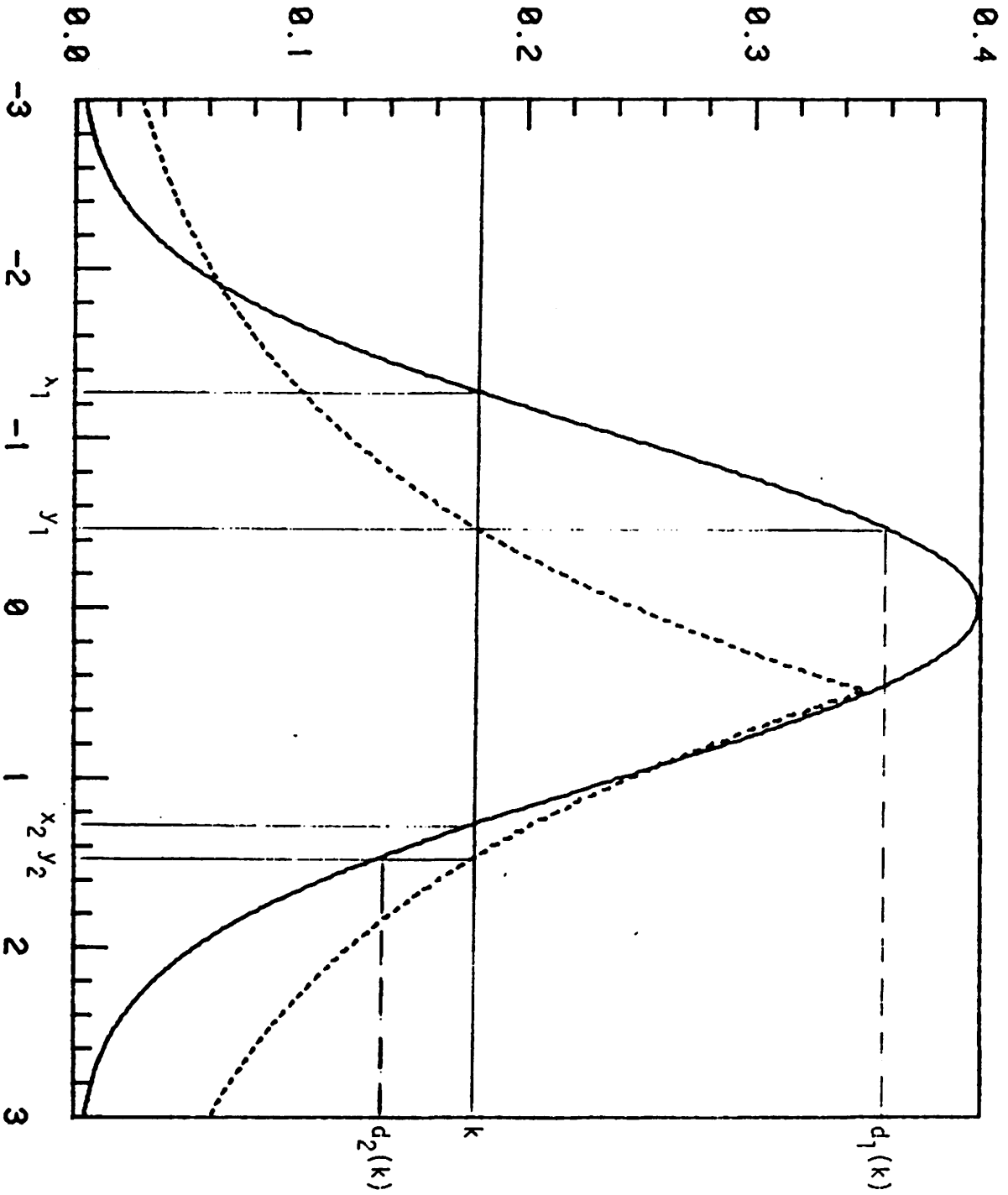$$p_0(k) \equiv p_0(k,h) = Pr_f(f(X) \geq k),$$

$$p_1(k,h) = Pr_f(f(X) \geq d_1(k,h)), \text{ and}$$

$$p_2(k,h) = Pr_f(f(X) \geq d_2(k,h)).$$

which are to be used as were p_0, p_1, and p_2 in the case $p=1$.

The method is completed by finding the directions h that minimize $p_1(k,h)$--

11

Fig. 3.1--Illustration of the contours method in one
dimension. Solid line = f, dotted line = g.

equivalently, maximize $d_1(k,h)$ over h to give $d_1(k) = \max d_1(k,h)$--and maximize $p_2(k,h)$--equivalently, minimize $d_2(k,h)$ over h to give $d_2(k) = \min d_2(k,h)$. If these extreme values of p_1 and p_2 differ greatly from $p_0(k)$, then f and g differ, with the interpretation of "greatly" again depending on the user's intent.

To recapitulate, the "contours" method of comparison, as this method will be called henceforth, consists of searching along a level curve of g for the directions h from f's mode in which f is farthest below and farthest above g. The scale used to measure these differences is derived using the density f. In section 4, this method is adapted for comparing likelihoods.

3.2 Integrals

If $f(x) \approx g(x)$ for all $x \in \Omega$, then for any function $\ell(x)$ and any curve segment S, the line integral

$$\int_S \ell(x)g(x)dx \quad \text{will be close to} \quad \int_S \ell(x)f(x)dx.$$

I will assume that f is the density of X, a p-dimensional random variable distributed as $N(\mu,\Sigma)$ for Σ nonsingular, and that g has the properties listed at the beginning of the section. Then the function ℓ and curve segment S to be used will give

$$I(f) \equiv \frac{2\pi^{p/2}|\Sigma|^{1/2}}{\Gamma(p/2)} \int_a^b r^{p-1}f(\mu + r\Sigma^{1/2}h)dr,$$

for h a unit length direction vector and a and b supplied by the user, and

$$I(g,h) = \frac{2\pi^{p/2}|\Sigma|^{1/2}}{\Gamma(p/2)} \int_a^b r^{p-1} g(\mu + r\Sigma^{1/2}h)dr$$

(the dependence of $I(f)$ and $I(g,h)$ on a and b will be suppressed for notational convenience). The motivation and calibration of this choice follow.

Let the scalars $b > a \geq 0$ be given, and define the ellipsoids

$$E_b = \{x | (x - \mu)'\Sigma^{-1}(x - \mu) \leq b^2\}$$

$$= \{x | x = \mu + r\Sigma^{1/2}h, \; \|h\| = 1, \; 0 \leq r \leq b\},$$

and E_a. Then

$$I(f) \equiv Pr_f(X \in E_b) - Pr_f(X \in E_a)$$

$$= Pr_f\left((X - \mu)'\Sigma^{-1}(X - \mu) \leq b^2\right) - Pr_f\left((X - \mu)'\Sigma^{-1}(X - \mu) \leq a^2\right)$$

$$= \int_a^b \frac{r^{p-1}e^{-r^2/2}}{\Gamma(p/2) \; 2^{p/2-1}} \; dr, \quad \text{for} \quad r = \left((x - \mu)'\Sigma^{-1}(x - \mu)\right)^{1/2}$$

$$= \frac{2\pi^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}{\Gamma(p/2)} \int_a^b \frac{|\Sigma|^{\frac{-1}{2}}}{(2\pi)^{\frac{p}{2}}} r^{p-1} \exp\{-([\mu + r\Sigma^{\frac{1}{2}}h]-\mu)'\Sigma^{-1}([\mu + r\Sigma^{\frac{1}{2}}h]-\mu)/2\}dr$$

for h an arbitrary direction vector

$$= \frac{2\pi^{p/2}|\Sigma|^{1/2}}{\Gamma(p/2)} \int_a^b r^{p-1} f(\mu + r\Sigma^{1/2}h)dr.$$

If $f(x) \approx g(x)$ for $x \in \Omega$, then

$$I(g,h) = \frac{2\pi^{p/2}|\Sigma|^{1/2}}{\Gamma(p/2)} \int_a^b r^{p-1} g(\mu + r\Sigma^{1/2}h)dr = I(f)$$

for arbitrary direction vectors h. By the contrapositive, if there is a direction h for which $I(g,h)$ and $I(f)$ differ "substantially," then $g(x)$ and $f(x)$ may be said to differ. Since $I(f)$ does not depend on h, the suggested usage for $I(g,h)$ is to find h^* and h_* that maximize and minimize, respectively, $I(g,h)$, and compare them to $I(f)$.

To interpret $I(g,h^*)$, define a function g^* as follows. For each fixed value of r in the half-open interval $(a,b]$, with h varying over the unit length direction vectors, define $g^*(\mu + r\Sigma^{1/2}h) = g(\mu + r\Sigma^{1/2}h^*)$. For $r \leq a$ or $r > b$, define g^* arbitrarily. Then for $a < r \leq b$, g^* has ellipsoidal contours: for each fixed value of r, $g^*(\mu + r\Sigma^{1/2}h = \xi(r)$ $(= g(\mu + r\Sigma^{1/2}h))$, so that as h varies with r fixed, the points $\mu + r\Sigma^{1/2}h$ trace out an ellipsoid.

Without loss of generality, let $\mu = 0$. Now if the variable of integration x varies over R^p,

$$\int_{x\epsilon E_b - E_a} g^*(x)\, dx = |\Sigma|^{1/2} \int_{a^2 < y'y \leq b^2} g^*(\Sigma^{1/2}y)\, dy, \text{ for } y = \Sigma^{-1/2}x, \text{ and}$$

$E_b - E_a$ the intersection of E_b and the complement of E_a. Now transform y to polar coordinates, letting $y = rh(\theta)$, for $\theta = (\theta_1, \ldots, \theta_{p-1})'$, where $r = (y'y)^{1/2} > 0$ and $h_i(\theta) = \cos \theta_i (\sin \theta_0 \sin \theta_1 \ldots \sin \theta_{i-1})$, $\sin \theta_0 = \cos \theta_p = 1$, $0 \leq \theta_j \leq \pi$, $j = 1,2,\ldots,p-2$, and $0 \leq \theta_{p-1} \leq 2\pi$. The Jacobian of the transformation from (r,θ) to y is $J = r^{p-1}(\sin^{p-2}\theta_1 \sin^{p-3}\theta_2 \ldots \sin \theta_{p-2})$ (Mardia, Kent, and

14

Bibby, 1979, pp. 35-6), so the above integral equals

$$|\Sigma|^{1/2} \int_{\{(r,\theta)|a<r\leq b\}} g^*(r\Sigma^{1/2}h(\theta)) \; J \; dr d\theta$$

$$= |\Sigma|^{1/2} \int_{\{(r,\theta)|a<r\leq b\}} g(r\Sigma^{1/2}h^*) \; J \; dr d\theta \quad \text{by the definition of } g^*$$

$$= |\Sigma|^{1/2} \int \frac{J}{r^{p-1}} \; d\theta \int_{a<r\leq b} r^{p-1} g(r\Sigma^{1/2}h^*) \; dr .$$

But $\int J/r^{p-1} \; d\theta$ (the integrand is independent of r) is the surface area of the unit spheroid of dimension p, which is $2\pi^{p/2}/\Gamma(p/2)$. Thus

$$\int_{x\in E_b'-E_a} g^*(x) \; dx = \frac{2\pi^{p/2}|\Sigma|^{1/2}}{\Gamma(p/2)} \int_a^b r^{p-1} g(r\Sigma^{1/2}h^*) \; dr = I(g,h^*) .$$

That is, if g^* is formed with the same shape in all directions (expanded and contracted by $\Sigma^{1/2}$) as g has in the direction h^*, then $I(g,h^*)$ is the area under g^* over the region $E_b - E_a$. By an exactly analogous argument, if g_* is formed having the same shape in all directions (expanded and contracted by $\Sigma^{1/2}$) as g has in the direction h_*, then $I(g,h_*)$ is the area under g_* over the region $E_b - E_a$.

Further, $I(g,h_*) \leq \text{Pr}_g(E_b - E_a) \leq I(g,h_*)$. To show the second inequality, let $\mu = 0$ again without loss of generality. Then $\text{Pr}_g(E_b-E_a)$

$$= \int_{a<\|\Sigma^{-1/2}x\|\leq b} g(x) \; dx = |\Sigma|^{1/2} \int_{a^2\leq y'y\leq b^2} g(\Sigma^{1/2}y) \; dy \quad \text{for } y = \Sigma^{-1/2}x$$

$$= |\Sigma|^{1/2} \int_{a<r\leq b} J \; g(r\Sigma^{1/2}h(\theta)) \; dr \; d\theta \quad \text{for } y = rh(\theta) \text{ as above}$$

15

$$= |\Sigma|^{1/2} \int \frac{J}{r^{p-1}} \left\{ \int_{a < r \le b} r^{p-1} g(r\Sigma^{1/2}h(\theta)) \, dr \right\} d\theta$$

$$\le |\Sigma|^{1/2} \int \frac{J}{r^{p-1}} \left\{ \int_a^b r^{p-1} g(r\Sigma^{1/2}h^*) \, dr \right\} d\theta \quad \text{by definition}$$

$$= I(g,h^*).$$

The first inequality follows from the analogous argument.

Using $I(g,h^*)$ and $I(g,h_*)$, then, amounts to using the shape of g along rays from f's mode μ in two directions, $\Sigma^{1/2}h^*/\|\Sigma^{1/2}h^*\|$ and $\Sigma^{1/2}h_*/\|\Sigma^{1/2}h_*\|$, to construct two surfaces g^* and g_* that have ellipsoidal contours and that satisfy

$$\int_{x \in E_b' - E_a} g^*(x) \, dx \ge \int_{x \in E_b' - E_a} g(x) \, dx \ge \int_{x \in E_b' - E_a} g_*(x) \, dx, \quad \text{with the}$$

first and last integrals being maximized and minimized respectively subject to the construction.

The purpose of defining $I(f)$ and $I(g,h)$ this way is to provide numbers that can be understood as probabilities. For example, it is easy to understand that if $I(g,h^*) - I(f) = 0.001 = I(f) - I(g,h_*)$ for all pairs (a,b), g and f are identical for any practical purpose. The main drawback with these definitions is that while $I(f)$ is a probability, $I(g,h^*) \ge I(f)$ is not, and may exceed unity. Nonetheless, because we are interested in either the difference $I(g,h^*) - I(f)$ or the ratio $I(g,h^*)/I(f)$, interpretability should not be impaired.

To recapitulate, the "integrals" method of comparison, as this method will

be called henceforth, consists of searching the directions h for those in which the integral

$$\int_a^b r^{p-1} g(\mu + r\Sigma^{1/2}h)dr$$

is large or small compared to

$$\int_a^b r^{p-1} f(\mu + r\Sigma^{1/2}h)dr.$$

These integrals are multiplied by the constant $2\pi^{p/2}|\Sigma|^{1/2}/\Gamma(p/2)$ to permit the latter to be interpreted as a probability.

3.3 Computing the assessment methods

The methods of sections 3.1 and 3.2 were defined without making approximations, in accordance with the requirements listed in section 2.3, but some approximations must be made to compute the two methods (that is, apart from the inevitable approximation that numerical optimization entails). As described below, our implementation of these methods (Hodges, 1985, Appendix) approximates the objective functions for the optimization problems that need to be solved to evaluate the measures for the two methods. It might appear that we have created a distinction without a difference: Bates and Watts use an approximation in defining their curvature measures and are able to compute them exactly; we define our measures without approximation but require approximations to compute them. However, as noted in section 2.2, there is a difference, in that improving on the approximation used for defining the curvatures would require a completely new method, while improving on the approximation used for computing

17

the integrals and contours methods only requires better computing technique.

In the remainder of this section, the approximations used for computing the contours and integrals methods are described.

For the definition of the objective functions to be optimized in computing the two methods of section 3, it will be assumed that g is the exact density of a random variable X, and f an approximation to g, where f is a $N(\mu, \Sigma)$ density, for Σ nonsingular. It is convenient to work with a transform of X, $Z = \Sigma^{-1/2}(X - \mu)$. The densities derived from g and f using this transformation will be called g_z and f_z respectively. The latter has the origin as its mode and circular level curves. The direction vectors of interest to the user are $h_\Sigma = \Sigma^{1/2}h/(h'\Sigma h)^{1/2}$, where h is the unit length direction vector found by the search method applied to f_z and g_z.

Transforming to f_z and g_z effects a considerable simplification in the objective function for the contours method. In terms of the original variable X, the desired level curve of g is

$$k = g(x) = g(\mu + \Sigma^{1/2}z); \qquad (3.1)$$

in terms of z, it is

$$|\Sigma|^{1/2}k = |\Sigma|^{1/2}g(\mu + \Sigma^{1/2}z) = g_z(z). \qquad (3.2)$$

If x satisfies (3.1), then $z = \Sigma^{-1/2}(x - \mu)$ satisfies (3.2), and conversely. For h a unit length direction vector, let $r_h > 0$ solve

$$g_z(r_h h) = |\Sigma|^{1/2} k. \qquad\qquad (3.3)$$

(This solution may not exist, or it may not be unique. These possibilities are discussed below.) If h^* maximizes

$$g_z(r_h h) - f_z(r_h h) = |\Sigma|^{1/2} k - f_z(r_h h), \qquad (3.4)$$

then h^* also maximizes the following two expressions:

$$|\Sigma|^{1/2} k - |\Sigma|^{1/2} f(\mu + r_h \Sigma^{1/2} h), \qquad\qquad (3.5)$$

which is just (3.4) rewritten, and

$$k - f(\mu + r_h \Sigma^{1/2} h); \qquad\qquad (3.6)$$

that is, h^* minimizes $f(\mu + r_h \Sigma^{1/2} h)$, and thus mimimizes $f(x)$ over $x \in C(g,k)$. But this means that $d_2(k, h_\Sigma)$ is minimized, or equivalently, that $p_2(k, h_\Sigma)$ is maximized, where $h_\Sigma = \Sigma^{1/2} h / (h' \Sigma h)^{1/2}$, as above.

Thus, if h^* maximizes (3.4), it also maximizes the original maximand $p_2(k, h_\Sigma)$; but since $f_z(r_h h) = (2\pi)^{-p/2} \exp(-r_h^2/2)$, h^* also maximizes r_h. So transforming to f_z and g_z allows us to maximize p_2 by finding the direction h in which the level curve (3.2) of g_z is farthest from the origin. By an exactly analogous argument, $p_1(k, h_\Sigma)$ is minimized by the direction h for which the level curve (3.2) of g_z is nearest to the origin.

The objective function used in the search, then is $S_c(h) = r_h$, for $r_h > 0$

19

solving (3.3). In developing the contours method in section 3.1, positive and negative values of r_h in (3.3) were allowed. Only positive values of r_h are allowed here but, as just shown, the resulting maximum and minimum are the same.

These last few paragraphs have two other implications. If f has circular contours, then maximizing $p_2(k,h)$ and minimizing $p_1(k,h)$ are equivalent to maximizing and minimizing, respectively, the distance from f's mode to g's level curve $g(x) = k$. If f has elliptical contours of the form $(x - \mu)'\Sigma^{-1}(x - \mu) = c$, then maximizing $p_2(k,h)$ and minimizing $p_1(k,h)$ are equivalent to maximizing and minimizing, repectively, the distance from μ to g's level curve, where the distance measure is defined by the inner product $(w,y) = w'\Sigma^{-1}y$.

The use of the objective function $S_c(h) = r_h$ can cause problems depending on how it is evaluated. If it is found by using a simple root finder in the direction h (as in Hodges, 1985), then if g_z's mode is placed in such a way that for some h there are two positive values of r satisfying (3.3), the root finder will detect that there are zero roots or an even number of roots and stop. This difficulty can also occur for densities with banana-shaped contours. The search routine can be written so that when this occurs, the occurence is pointed out to the user, but never unambiguously, because other features of the exact density can cause the simple root finder to think that there is an even number of roots in the direction h.

The objective function for the integrals method does not reap such a large gain in simplicity from transforming X to Z, though the computations are a bit simpler. Thus

$$I(g,h) = \frac{2\pi^{p/2}|\Sigma|^{1/2}}{\Gamma(p/2)} \int_a^b r^{p-1} g(\mu + r\Sigma^{1/2}h)dr$$

$$= \frac{2\pi^{p/2}}{\Gamma(p/2)} \int_a^b r^{p-1} g_z(rh)dr.$$

It would be ideal to use as an objective function the exact integral $I(g,h)$, but the state of numerical integration is such that this is not practical, so an approximation to the ideal is used in the search. I have used the trapezoidal approximation

$$I(g,h) \approx S_i(h) = \frac{2\pi^{p/2}}{\Gamma(p/2)} \left(\frac{1}{2}(b-a)\{b^{p-1}g_z(bh) + a^{p-1}g_z(ah)\}\right)$$

$$= \frac{\pi^{p/2}}{\Gamma(p/2)} (b-a)\{b^{p-1}g_z(bh) + a^{p-1}g_z(ah)\}.$$

Clearly, it would be possible to use a finer approximation to $I(g,h)$, e.g. two trapezoids or a 3-point Simpson's rule, but I have found this single trapezoid satisfactory and quite fast in test runs. The trapezoid assumption will do poorly when the mode of g_z lies on h between ah and bh. This possibility can be obviated by performing the search for more than one pair (a,b). Once the search has converged to a direction, $I(g,h)$ can be evaluated numerically, so that the values of $I(g,h^*)$ and $I(g,h_*)$ should be quite accurate for the directions h^* and h_*, though h^* and h_* will not in general be exactly the optimal directions.

In the version of these methods in Hodges (1985), the search method used is a gradient projection algorithm (Luenberger, 1984, sections 11.4 and 11.5), with all derivatives taken numerically. This implementation uses two stopping

conditions. If h_j is the current trial value of the direction h, $\nabla S(h_j)$ is the gradient of the relevant objective function at the trial value h_j, and ϵ is a small number supplied by the user, the two stopping conditions are: (i) $h_j'h_{j+1}$ > 1 - ϵ, i.e. the trial value does not change much, and (ii) $|h_j'\nabla S(h_j)|$ > 1 - ϵ, which means that the first order necessary condition for a Lagrange multiplier solution to this optimization problem is approximately satisfied. Finally, the program in Hodges (1985) allows the user to supply starting values or to call an IMSL subroutine that generates starting values.

At this point, a word is in order about the assumption that g is unimodal. In many cases, g will have either multiple modes or an unknown number of modes. The difficulty this presents for the integrals method is essentially the same as the difficulty presented by an off-center single mode of g. The remedy suggested for that problem will be helpful here as well. It is reasonable to expect that if the simple trapezoid approximation misses a mode of g, some other direction will provide evidence of g's divergence from f.

For the contours method, the difficulty introduced by multiple modes is also similar to that created by an off-center mode and can be handled similarly.

4. EXAMPLES, AND AN EXTENSION TO APPROXIMATIONS TO LIKELIHOODS

4.1 Possible uses for the two methods

Because the two methods in section 3 are simply methods for comparing two functions, they are applicable in a wide range of statistical situations. The integrals method is useful for examining the normal approximation to the posterior density. If one would like to use a central region of probability $1 - \alpha$, the constants a and b can be set to 0 and $\sqrt{\chi^2_{p,\alpha}}$, respectively, where $Pr(\chi^2_p \leq \chi^2_{p,\alpha}) = 1 - \alpha$, and bounds on the actual probability of the central region can be calculated as in section 3.2.

Hodges (1985, section 3.4) contains an example in which the contours method is used to examine the appropriateness of forming an approximate confidence region by using the usual normal approximation to the distribution of the m.l.e. The integrals method can also be used (Hodges, 1985, section 3.5a) to study the dependence on sample size and dimension of the accuracy of an approximation used by Johnson and Geisser (1982, 1983) for computing the predictive influence function (PIF's) of subsets of points in linear regression models.

In section 4.2, the contours method will be extended to allow comparison of likelihood regions and applied to the Fieller-Creasy problem (as presented in Cook and Witmer, 1985), and in section 4.3, the integrals method is applied to examining two approximations of Geisser (1970) and Lee and Geisser (1972) to the predictive density for growth curve problems.

4.2 Likelihood regions

Under regularity conditions, if θ° is the true value of θ,

$$W(\theta^0) = -2 \log\left(\frac{L(\theta^0;y)}{L(\hat{\theta};y)}\right) \qquad (4.1)$$

is approximately distributed as χ_p^2 (cf. Cox and Hinkley, 1974, pg. 322), so that $W(\theta)$ is an approximate pivotal and it can be used to form approximate confidence regions. An approximate $100(1-\alpha)\%$ confidence region would be the region in parameter space satisfying

$$1-\alpha \doteq \Pr(W(\theta) \leq c) \qquad (4.2)$$

for c the appropriate quantile of χ_p^2. But

$$\{\theta | W(\theta) \leq c\} = \{\theta | \log L(\theta;y) \geq c'\} \qquad (4.3)$$

i.e., the level curves of the likelihood have an approximate confidence interpretation.

Finding the region $\{\theta | \log L(\theta;y) \geq c'\}$ can be computationally impractical, and $\log L(\theta;y)$ is commonly approximated by the second order Taylor series around $\hat{\theta}$,

$$\log L(\theta;y) \doteq \log L(\hat{\theta};y) - (\theta - \hat{\theta})'I_\theta(\theta - \hat{\theta})/2,$$

where

$$I_\theta = \left(\left(-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log L(\theta;y)\bigg|_{\theta = \hat{\theta}} \right)\right)_{p\times p}$$

is the observed Fisher information. When this is inserted into the left hand side of the inequality in (4.3), it can be expressed as

$$1 - \alpha \doteq \Pr((\theta - \hat{\theta})'I_\theta(\theta - \hat{\theta}) \leq c''), \qquad (4.4)$$

and

$$\{\theta | (\theta - \hat{\theta})' I_\theta (\theta - \hat{\theta}) \le c\} = \{\theta | f(\theta) \ge \frac{|I_\theta|^{1/2} e^{-c/2}}{(2\pi)^{p/2}}\},$$

where $f(\theta)$ is the density of Θ, a $N(\hat{\theta}, I_\theta^{-1})$ random variable. Note that the nominal confidence coefficient $1 - \alpha$ is exactly

$$Pr_f((\Theta - \hat{\theta})' I_\theta (\Theta - \hat{\theta}) \le c) = Pr_f(f(\Theta) \ge \frac{|I_\theta|^{1/2} e^{-c/2}}{(2\pi)^{p/2}}).$$

Similarly,

$$\{\theta | -2 \log[\frac{L(\theta;y)}{L(\hat{\theta};y)}] \le c\} = \{\theta | g(\theta) \ge \frac{|I_\theta|^{1/2} e^{-c/2}}{(2\pi)^{p/2}}\} \qquad (4.5)$$

$$\text{for} \quad g(\theta) = \frac{|I_\theta|^{1/2} L(\theta;y)}{(2\pi)^{p/2} L(\hat{\theta};y)}.$$

With this manipulation, we are almost in the situation of section 3: we are interested in the level curves of $g(\theta)$ and we would like to approximate them with the level curves of $f(\theta)$, a normal density. This is not exactly the situation of section 3, because g will be a probability density only by coincidence. Nonetheless, the contours method can be applied directly to $g(\theta)$ and $f(\theta)$ with a change in the interpretation of the results, as will now be shown.

Define the function

$$k(c) = \frac{|I_\theta|^{1/2} e^{-c/2}}{(2\pi)^{p/2}} , \text{ with inverse function}$$

$$c(k) = -2 \log((2\pi)^{p/2} k |I_\theta|^{-1/2}).$$

Let θ_0 be a point on the contour $g(\theta) = k(c)$ of g. The contour of f which passes through θ_0 is $f(\theta) = f(\theta_0)$, and from above this contour is the normal approximate confidence region with nominal confidence coefficient

$$Pr_f(f(\theta) \geq f(\theta_0)) = Pr_f\{(\theta - \hat{\theta})'I_{\hat{\theta}}(\theta - \hat{\theta}) \leq c(f(\theta_0))\}$$

$$= Pr\{\chi_p^2 \leq c(f(\theta_0))\}.$$

Thus, if we apply the contours method to $f(\theta)$ and $g(\theta)$ as if they were both probability densities, the minimum value of $p_1(k,h)$ obtained will be the nominal confidence coefficient of the largest normal approximate region entirely contained in the exact region (4.5), and the maximum value of $p_2(k,h)$ obtained will be the nominal confidence coefficient of the smallest normal approximate region which contains the exact region (4.5). If the expected Fisher information

$$I_\theta^e = ((\{E \frac{-\partial^2}{\partial\theta_i \partial\theta_j} \log L(\theta;y)\}\bigg|_{\theta = \hat{\theta}}))_{p\times p}$$

were used in place of $I_{\hat{\theta}}$ in (4.4) and the subsequent argument, the same interpretation will hold for approximate regions formed using I_θ^e.

To understand what drives the contours method in this application, let $f^*(\theta) = (\theta - \hat{\theta})'I_{\hat{\theta}}(\theta - \hat{\theta})$ and $g^*(\theta) = -2[\log L(\theta;y) - \log L(\hat{\theta};y)]$, so that the approximate and exact likelihood regions are $\{\theta|f^*(\theta) \leq c\}$ and $\{\theta|g^*(\theta) \leq c\}$ respectively. The boundaries of the two regions are

$$c = f^*(\theta) = (\theta - \hat{\theta})'I_\theta(\theta - \hat{\theta})$$

$$c = g^*(\theta) \approx (\theta - \hat{\theta})'I_\theta(\theta - \hat{\theta}) + \{\text{third degree terms}\}.$$

If the approximate region differs from the exact region, to the first order of approximation the difference is caused by the third degree terms in the expansion of g^*. This is essentially the measure used by Jennings (1982) to characterize the adequacy of the normal approximation to the likelihood region for logistic regression models.

If the approximate likelihood region is formed using I_θ^e instead of I_θ, the difference between $f^*(\theta)$ and $g^*(\theta)$ contains, to the first order of approximation, a term $(\theta - \hat{\theta})'(I_\theta - I_\theta^e)(\theta - \hat{\theta})$ as well as the third degree terms.

The interpretation of the integrals comparison cannot be adapted so neatly. Still, because $f(\theta)$ is being used to approximate $g(\theta)$, and $f(\theta)$ is a density, the integrals comparison can be applied and $I(g,h)$ interpreted as if $g(\theta)$ were also a density. Although with likelihood methods we are not interested in probabilities of sets, the integrals method still provides information about the shape of g relative to f.

This re-interpretation of the contours method for use with likelihood regions extends to subsets of the parameter vector. The development is identical, except that $L(\theta;y)/L(\hat{\theta};y)$ is replaced by

$$\max_{\theta_2} \frac{L(\theta_1,\theta_2;y)}{L(\hat{\theta};y)} \quad \text{and} \quad \chi_p^2 \quad \text{by} \quad \chi_q^2.$$

To give an example of this usage of the contours method, consider the nonlinear regression problem as defined in section 2.2, and assume σ^2 is known.

Then the exact likelihood region (4.3) is

$$\left\{ \theta \mid \frac{S(\theta) - S(\hat{\theta})}{\sigma^2} \le c \right\} \qquad (4.6)$$

where $S(\theta) = \sum_{i=1}^{n} (y_i - F(x_i;\theta))^2$ and c satisfies $Pr(\chi_p^2 \le c) =$

1 - α. This exact region is commonly approximated by replacing $(S(\theta) - S(\hat{\theta}))$

with $(\theta - \hat{\theta})'V.'V.(\theta - \hat{\theta})$, where V. is the n×p matrix with (i,j) element

$\partial F(x_i;\theta)/\partial\theta_j$ evaluated at $\hat{\theta}$. This tangent plane approximation is used in

defining the calibration of the curvature measures in section 2.2. Since σ^2 is

assumed known, $(\theta - \hat{\theta})'V.'V.(\theta - \hat{\theta})/\sigma^2$ is approximately χ_p^2, by the tangent plane

approximation, and the region

$$\left\{ \theta \mid (\theta - \hat{\theta})'V.'V.(\theta - \hat{\theta}) \le \sigma^2 c \right\} \qquad (4.7)$$

approximates (4.6).

We can now apply the methods of section 3 to assessing the adequacy of this

approximation, with f(θ) the density of a $N(\hat{\theta},\sigma^2(V.'V.)^{-1})$ random variable, and

$$g(\theta) = \frac{|V.'V.|^{1/2}}{(2\pi\sigma^2)^{p/2}} \exp\left\{ \frac{-(S(\theta) - S(\hat{\theta}))}{2\sigma^2} \right\} . \qquad (4.8)$$

As an example, consider the Fieller-Creasy problem. The treatment here follows

Cook and Witmer (1985). Suppose that y_i is an observable normal random variable

with known variance σ^2 and mean $\theta_1 x_i + \theta_1\theta_2(1 - x_i)$, where $x_i = 1$ for population

1 and $x_i = 0$ for population 2. Then population 1 has unknown mean θ_1,

population 2 has unknown mean $\theta_1\theta_2$, and the ratio of population 2's mean to

population 1's mean is θ_2. Assume that $y_1,...,y_n$ are observations on population

1 and $y_{n+1}, \ldots, y_{2n}$ are observations on population 2.

The likelihood function is proportional to

$$\exp(\frac{-n}{2\sigma^2} \{(\theta_1 - \hat{\theta}_1)^2 + (\theta_1\theta_2 - \hat{\theta}_1\hat{\theta}_2)^2\})$$

where the equations

$$\hat{\theta}_1 = \sum_{i=1}^{n} y_i/n \quad \text{and} \quad \hat{\theta}_1\hat{\theta}_2 = \sum_{i=n+1}^{2n} y_i/n$$

give the m.l.e.'s for θ_1 and θ_2. These m.l.e.'s are sufficient statistics for this problem. Assume that $(\hat{\theta}_1, \hat{\theta}_2) = (3,0)$ was observed and that $\sigma^2/n = 6/5.99147$. Then for $\alpha = 0.05$ and $c = 5.99147$, the exact likelihood region is

$$\{\theta | g(\theta) \leq \frac{5.99147}{4\pi} \exp(\frac{-(5.99147)^2}{12})\} \qquad (4.9)$$

-- as Cook and Witmer (1985) show, this likelihood region happens to have exact coverage of $1 - \alpha$ -- and the approximate region is

$$\{\theta | f(\theta) \leq \frac{5.99147}{4\pi} \exp(\frac{-(5.99147)^2}{12})\}. \qquad (4.10)$$

These contours are presented in Figure 4.1. The ellipse is the normal approximation and the wedge is the exact region. The results of the two comparison methods are in Table 4.1, where a and b were chosen so that $Pr_f(E_b) = 0.99$, $Pr_f(E_a) = 0.95$, and $I(f) = 0.04$, and k was chosen to give $p_0(k) = 0.95$.

By construction, the intrinsic curvature of the solution locus $\eta(\theta) = ((\theta_1 x_i + \theta_1\theta_2(1 - x_i)))_{n \times 1}$ is identically zero. From Cook and Witmer's formula, the parameter effects curvature is 0.33, and the guide value is $(5.991)^{-1/2} =$

0.41 (because σ^2 is assumed known, we compare to χ_p^2 instead of the F

distribution on p and ν degrees of freedom). Bates and Watts' criterion

indicates that the approximation shown in Figure 4.1 should be reasonable, yet

apparently it is not.

The two comparison methods of section 3 clearly indicate the inadequacy of

the elliptical region as an approximation to the wedge-shaped exact region.   In

directions 3 and 7, the exact region extends

Table 4.1

Number	direction h	$I(f) - I(g,h)$	$I(g,h)/I(f)$	$p_1(k,h)$ and $p_2(k,h)$
1	(-1.000, 0.000)	0.000	1.000	0.9501*
2	(-0.870, 0.494)	-0.236	6.889	---
3	(-0.826, 0.564)	---	---	>0.99995
4	(0.916, 0.400)	0.0380	0.0501	0.8476
5	(1.000, 0.000)	0.000	1.000	0.9501
6	(0.916,-0.400)	0.0380	0.0501	0.8476
7	(-0.826,-0.564)	---	---	>0.99995
8	(-0.870,-0.494)	-0.236	6.889	---

*The symbol "---" means that the direction was not a local maximum or minimum
of the objective function at the top of the column.

out past the edge of the ellipsoid with nominal confidence coefficient 0.99995.

In directions 4 and 6, the exact region reaches to the edge of the ellipsoid of

nominal confidence coefficient 0.8476. The integrals method corroborates this.

However, in view of Weiss' (1985) simulations for this problem, the apparent

superiority of the contours and integrals method for this problem must be

considered qualified.

This adapted version of the contours method presented in this section can

be extended to subsets of the parameter vector, although computation of the

method is very difficult except for single parameter subsets (Hodges, 1985,

section 3.3c).

Fig. 4.1 -- Contours of the likelihoods for the
Fieller-Creasy problem
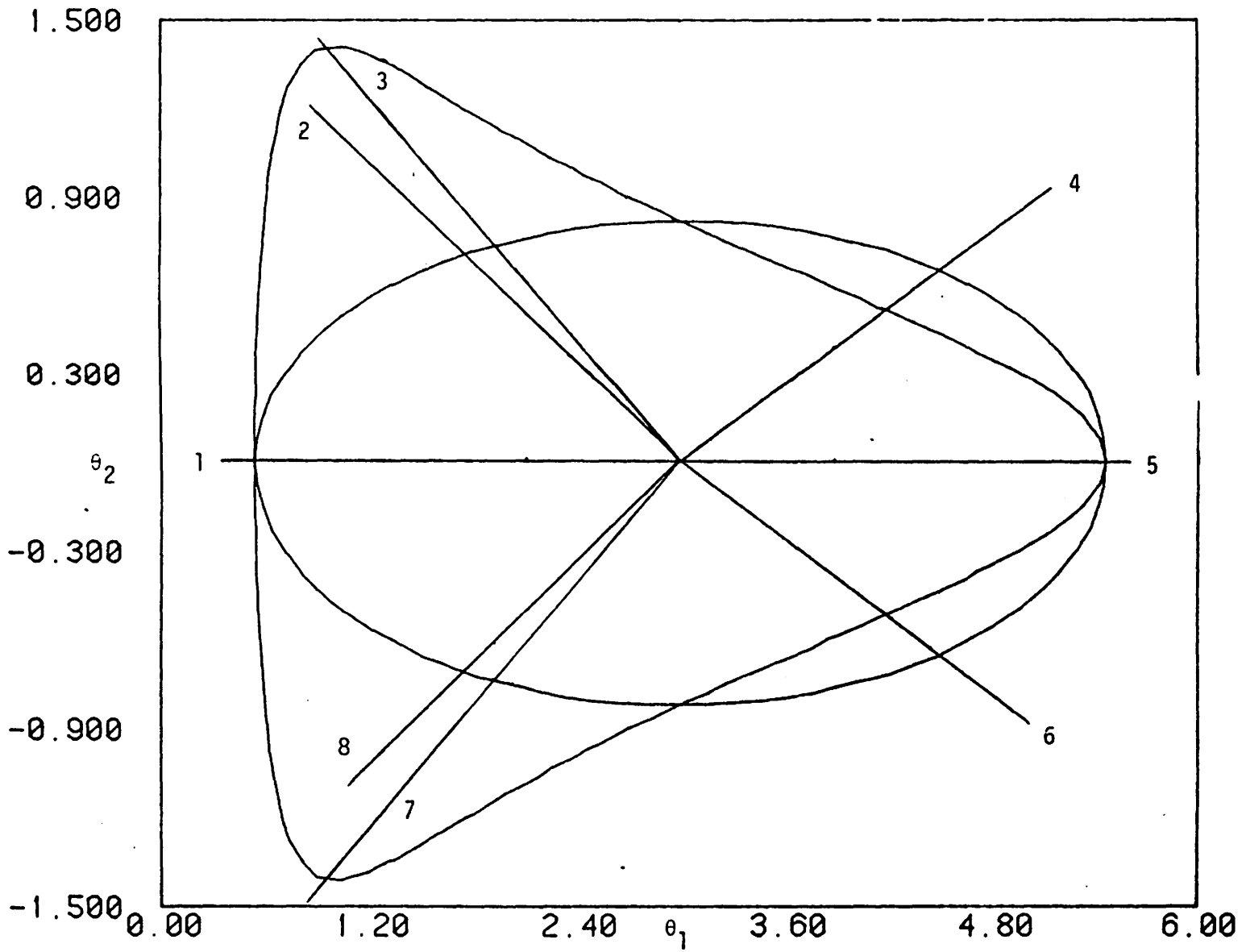
4.3 Predictive densities

This section contains a more fully developed example of how the accuracy assessment methods can be used to examine a particular approximation.

Geisser (1970) and Lee and Geisser (1972) developed a Bayesian analysis of growth curve models with particular interest in predictions from those models, and Lee and Geisser (1975) applied this analysis to the Potthoff-Roy data (given in Lee and Geisser, 1975). In these models the p×N matrix of observables Y is assumed to have expected value XτA, where X is a known p×m matrix of rank m < p, A is a known r×N matrix of rank r < N, and τ is an unknown m×r matrix. The columns of Y are assumed to be independent p-dimensional multivariate normal random variables with common unknown covariance matrix Σ. For the Potthoff-Roy data set, Lee and Geisser assumed that Σ had the simple form $\Sigma = X\Gamma X' + Z\theta Z'$ for Z a known p×(p-m) matrix of rank (p-m) satisfying Z'X = 0 and Γ and θ arbitrary m×m and (p-m)×(p-m) positive semi-definite matrices. This is called "Rao's simple structure," and a likelihood ratio test does not contraindicate its use for this data.

One prediction of interest is of K future p-dimensional vectors forming the p×K matrix V, which has expectation XτF for some appropriate known r×K matrix F. If the convenient prior density for $(\tau, \Gamma^{-1}, \theta^{-1})$

$$\pi(\Gamma^{-1}, \theta^{-1}, \tau) \propto |\Gamma|^{(m+1)/1} |\theta|^{(p-m+1)/2}$$

is used, the exact predictive density for V can be derived, but it is awkward to evaluate. However, its expectation and covariance matrix can be obtained, and Lee and Geisser (1972) propose approximating the exact predictive density of V

31

by a matrix normal density with expectation and covariance equal to those of the exact density.

Lee and Geisser also propose another approximation. If Γ and θ are assumed to be known, the predictive distribution of V is a normal distribution with a known mean and a covariance matrix that is a function of Γ and θ. The approximation is used by substituting estimates for the unknown Γ and θ, and Lee and Geisser suggest using unbiased estimates, m.l.e.'s, or posterior expectations. Here we will examine only the approximation obtained using the unbiased estimates.

The accuracy of these approximations can be checked using the methods of section 3; this will be exemplified with the Potthoff-Roy data.

These data consist of four dental measurements on each of eleven girls and sixteen boys. The postulated growth curve is linear, so that $p = 4$, $m = r = 2$, and $N = 27$, the design matrices X and A are

$$
X = \begin{bmatrix} 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \qquad A = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix}
$$

(where the first block of A has 11 columns and the second block has 16 columns), and τ is 2×2. If we wish to predict the four measurements for the next girl, K = 1 and F = (1,0)'. Then from Geisser (1970), the exact predictive density of V is

$$
g(V) = \frac{c_1 \, |DYY'D'|^{27/2} \, |BSB'|^{25/2}}{|DYY'D + DVV'D'|^{14} \, |BSB' + c_2(BV - T)(BV - T)'|^{13}}
$$

$$(4.11)$$

where $c_1 = 143/80\pi^2$, $c_2 = 11/12$, $T' = (22.648, 0.47955)$,

$$B = \frac{1}{20} \begin{bmatrix} 5 & 5 & 5 & 5 \\ -3 & -1 & 1 & 3 \end{bmatrix}, \quad D = \frac{1}{20} \begin{bmatrix} 5 & -5 & -5 & 5 \\ -1 & 3 & -3 & 1 \end{bmatrix} \quad \text{for } Z = \begin{bmatrix} 1 & -1 \\ -1 & 3 \\ -1 & -3 \\ 1 & 1 \end{bmatrix},$$

$$DYY'D' = \begin{bmatrix} 7.1719 & -0.32812 \\ -0.32812 & 3.1994 \end{bmatrix} \quad \text{with determinant } 22.838, \text{ and}$$

$$BSB' = \begin{bmatrix} 94.491 & 1.7008 \\ 1.7008 & 2.9586 \end{bmatrix} \quad \text{with determinant } 276.67 \text{ (interested}$$

readers are referred to Geisser, 1970, for details).

From Lee and Geisser (1972), expression (6.4), the exact predictive covariance of V is

$$\Sigma_V = \frac{12}{11} \left(\frac{1}{22} XBSB'X' \right) + \frac{1}{24} (ZDYY'D'Z')$$

$$= \begin{bmatrix} 5.9594 & 4.0349 & 4.2052 & 3.5307 \\ & 6.2242 & 3.6379 & 4.4879 \\ & & 6.4175 & 4.8190 \\ & & & 6.9167 \end{bmatrix},$$

and the exact predictive mean is $\mu = XTF = (21.209, 22.168, 23.127, 24.086)'$.
We will refer to this approximation, a 4-variate normal density with mean μ and covariance Σ_V, as (6.4).

Also, from Lee and Geisser (1972), expression (8.7), the second approximation (which uses the unbiased estimates of Γ and θ) has covariance matrix

$$\Sigma_u = \frac{12}{11} X(\frac{1}{25}BSB')X' + Z(\frac{1}{27}DYY'D')Z'$$

$$= \begin{bmatrix} 5.2484 & 3.5440 & 3.7017 & 3.1085 \\ & 5.5090 & 3.1934 & 3.9500 \\ & & 5.6600 & 4.2350 \\ & & & 6.0903 \end{bmatrix}$$

and mean μ. This approximation will be referred to as (8.7). Note that

dividing Σ_u elementwise by Σ_v gives these quotients:

$$\begin{bmatrix} 1.14 & 1.14 & 1.14 & 1.14 \\ & 1.13 & 1.14 & 1.14 \\ & & 1.13 & 1.14 \\ & & & 1.14 \end{bmatrix} ,$$

which means that we can expect substantial similarity between the directions of

greatest difference found in comparing them to (4.11).

The results of comparing (6.4) and (8.7) to (4.11) are in Tables 4.2 and

4.3 respectively. For each of these two comparisons, seven integrals and seven

contours were examined. Four of the integrals gave the probabilities of central

regions, i.e. ellipsoids having probabilities (under the respective normal

approximations) of 0.05, 0.5, 0.8, and 0.95. These integrals are numbered 1

through 4, respectively; the integral numbers correspond to the columns in

Table 4.2a and in the upper part of Table 4.3. The other three integrals were

between ellipsoids containing probability 0.48 and 0.52, 0.78 and 0.82, and 0.93

and 0.97 under the normal approximation; these integrals are numbered 5 through

7, respectively. These last three integrals allow us to examine the

approximations at the edges of the central regions containing probabilities 0.5,

0.8, and 0.95. The seven contours were the ellipsoids containing 0.025, 0.25,

0.4, 0.475, 0.5, 0.8, and 0.95 of the probability under the normal approximation and are numbered 1 to 7 respectively; these contour numbers correspond to the columns in Table 4.2b and in the lower part of Table 4.3. The directions giving local maxima and minima are almost identical for (6.4) and (8.7), and they did not vary much over the seven integrals and seven contours. Thus, the directions were numbered and used for both (6.4) and (8.7). In both Tables, the direction numbers correspond to the rows, and the directions themselves are listed at the bottom of Table 4.2b.

The four central regions, integrals 1-4, will be examined first. Integral #1 shows that (6.4) is flatter at the mode than the exact density (because $I(g,h)/I(f) > 1$), while (8.7) is more peaked than the exact density. For neither approximation do the ratios $I(g,h)/I(f)$ depend much on the direction, so we have a good assessment of the size of the approximation error. For (6.4), the actual probability content of the approximate central region is about 14% higher than the nominal probability of 0.05, while for (8.7) the actual probability content of the approximate central region is about 8% lower than the nominal probabilty of 0.05. This trend persists for integral #2, with approximation (6.4) giving a nominal probability for the central region that is about 3% to 6% low and approximation (8.7) giving a nominal probability for the central region that is about 7 1/2% to 10% high. For integral #3, there are some directions in which approximation (6.4) is too high, and some in which it is too low. These cancel each other out to some extent, although undoubtedly not perfectly. Approximation (8.7) is still too high in all directions, by about 3 1/2% to 8 1/2%. Finally, for integral #4, approximation (6.4)'s high and low points cancel each other out better than for integral #3. For

35

Table 4.2a

<center>integrals</center>

dir.	1	2	3	4	5	6	7
1	--	1.0501	1.0322	1.0444	0.9988	1.0317	1.2948
2	1.1365	--	--	--	--	--	--
3	--	1.0501	1.0330	1.0454	0.9988	1.0329	1.2965
4	1.1369	--	--	--	--	--	--
5	--	1.0298	0.9826	0.9625	0.9443	0.8658	0.8599
6	--	1.0304	0.9833	0.9632	0.9450	0.8666	0.8609
7	1.1405	1.0604	1.0438	1.0554	1.0127	1.0438	1.2917
8	1.1364*	1.0298	0.9826	0.9625	0.9443	0.8658	0.8599
9	--	1.0301	0.9828	0.9625	0.9445	0.8659	0.8601
10	1.1364*	--	--	--	--	--	--
11	--	1.0305	0.9832	0.9630	0.9450	0.8665	0.8608
12	1.1367*	--	--	--	--	--	--
13	1.1405*	1.0604	1.0438	1.0555	1.0128	1.0440	1.2919
14	1.1366*	1.0302	0.9830	0.9629	0.9447	0.8664	0.8606
15	1.1405*	--	--	--	--	--	--
16	1.1405*	--	--	--	--	--	--

	integrals		contours
number	$Pr(E_a)$	$Pr(E)$	$Pr_f(f \geq k)$
1	0.0001	0.05	0.025
2	0.0001	0.5	0.25
3	0.0001	0.8	0.4
4	0.0001	0.95	0.475
5	0.48	0.52	0.5
6	0.78	0.82	0.8
7	0.93	0.97	0.95

* = direction slightly different from that shown
-- = the direction was not a local maximum or minimum for this
  integral or contour

Table 4.2b

			contours				
dir.	1	2	3	4	5	6	7
1	--	0.2631	0.4034	0.4754	0.4997	0.8050	0.9612
2	0.0488	--	--	--	--	--	--
3	--	0.2634	0.4037	0.4756	0.5000	0.8052	0.9613
4	0.0489	--	--	--	--	--	--
5	--	0.2572	0.3916	0.4603	0.4835	0.7783	0.9432
6	--	0.2573	0.3918	0.4605	0.4837	0.7785	0.9432
7	0.0501	0.2668	0.4077	0.4796	0.5039	0.8068	0.9610
8	0.0487*	0.2572	0.3916	0.4603	0.4835	0.7783	0.9432
9	--	--	0.3917	0.4603	0.4836	0.7784	0.9432
10	0.0487*	0.2571	--	--	--	--	--
11	--	--	0.3918	0.4605	0.4837	0.7785	0.9432
12	0.0488*	0.2573	--	--	--	--	--
13	0.0501*	0.2669	0.4077	0.4796	0.5039	0.8068	0.9610
14	0.0488*	0.2573	0.3917	0.4605	0.4836	0.7784	0.9432
15	0.0501*	0.2668	--	--	--	--	--
16	0.0501*	0.2668	--	--	--	--	--

directions

1 (-0.357,-0.445,-0.534,-0.623)   9 (-0.150, 0.230,-0.704,-0.655)

2 (-0.301,-0.523,-0.534,-0.592)   10 (-0.133, 0.185,-0.696,-0.681)

3 (0.357, 0.445, 0.534, 0.623)   11 (0.150,-0.230, 0.704, 0.655)

4 (0.301, 0.523, 0.534, 0.592)   12 (0.131,-0.191, 0.696, 0.679)

5 (-0.194,-0.648,-0.521,-0.521)   13 (-0.451, 0.373, 0.616,-0.528)

6 (0.193, 0.648, 0.521, 0.521)   14 (0.551, 0.641, 0.531,-0.060)

7 (0.130,-0.564, 0.755,-0.308)   15 (-0.341, 0.816,-0.460, 0.078)

8 (-0.292,-0.613,-0.245,-0.692)   16 (0.393,-0.769, 0.503,-0.039)

for Table 4.3:

17 (0.066,-0.229,-0.523,-0.818)   18 (-0.048, 0.240, 0.526, 0.814)

Table 4.3

integrals

dir.	1	2	3	4	5	6	7
1	0.9179	0.9280	0.9645	1.0167	0.9622	1.1320	1.6404
3	0.9183	0.9286	0.9653	1.0176	0.9631	1.1333	1.6424
5	0.9171	0.9098	0.9192	0.9371	0.9142	0.9726	1.1566
6	0.9174	0.9102	0.9198	0.9377	0.9148	0.9735	1.1578
7	0.9186	0.9281	0.9622	1.0106	0.9601	1.1182	1.5862
8	0.9171	0.9098	0.9193	0.9372	0.9143	0.9728	1.1568
9	0.9172	0.9099	0.9193	0.9372	0.9144	0.9729	1.1569
11	0.9173	0.9102	0.9198	0.9377	0.9149	0.9736	1.1578
13	0.9186	0.9281	0.9621	1.0105	0.9600	1.1181	1.5860
14	0.9173	0.9101	0.9196	0.9375	0.9147	0.9733	1.1575
17	--	--	--	1.0168	--	--	1.6406
18	--	--	--	1.0175	--	--	1.6422

contours

dir.	1	2	3	4	5	6	7
1	0.0112	0.2194	0.3774	0.4592	0.4868	0.8212	0.9710
3	0.0113	0.2197	0.3777	0.4595	0.4871	0.8213	0.9711
5	0.0112	0.2142	0.3660	0.4444	0.4708	0.7955	0.9563
6	0.0112	0.2144	0.3662	0.4446	0.4710	0.7957	0.9563
7	0.0113	0.2197	0.3772	0.4586	0.4861	0.8189	0.9696
8	0.0112	0.2143	0.3660	0.4444	0.4708	0.7955	0.9563
9	0.0112	0.2143	0.3661	0.4444	0.4708	0.7956	0.9563
11	0.0112	0.2144	0.3662	0.4446	0.4710	0.7957	0.9563
13	0.0113	0.2196	0.3771	0.4586	0.4861	0.8189	0.9696
14	0.0112	0.2144	0.3662	0.4445	0.4709	0.7956	0.9563
17	--	--	--	0.4592	--	--	0.9710
18	--	--	--	0.4595	--	--	0.9711

approximation (8.7), some ratios are greater than unity, but for each direction (6.4) has a higher ratio $I(g,h)/I(f)$, so (8.7) will still overestimate the exact integral by a few percent.

In sum, (6.4) is better for the three larger regions, (8.7) is better for the smallest region, but in no case are the differences very large.

Integrals 1, 5, 6, and 7 help explain the trends in integrals 1-4. For (6.4) there are two groups of directions: in one group, the approximate integral exceeds the exact integral by an increasing amount (directions 5, 6, 8, 9, 11, 14), and in the other group, the exact integral exceeds the approximate integral by an increasing amount (directions 1, 3, 7, 13). These directions are the same ones in which the ratios for integral #4 are less than or greater than unity, respectively. Approximation (8.7) shows a similar pattern, except that for the first group of directions the exact integral also exceeds the approximate integral, though not by as much as for the second group of directions.

Integrals 5 through 7 indicate that both approximations have some deficiencies in their tails. In particular, if we wanted to make a statement about the probability **outside** some central region, integrals 5-7 indicate that (8.7) would underestimate that probability, because for integral 7 it underestimates in all directions, while (6.4) would be closer to the exact value because it would overestimate in some directions and underestimate in others. However, it is not clear from these seven integrals what would happen far out in the tails.

All of this is corroborated by the contours comparisons.

Both of these approximations do fairly well. Neither is a resounding

winner over the other, but (6.4) is probably a little bit better.

5. CONCLUSION

How do the methods of section 3 measure up against the desirable qualities listed in section 2.3? For the examples shown in section 4, the methods do use directly relevant information about the entities of interest, i.e. the likelihood function in section 4.2 and the predictive density in section 4.3. For some of the applications listed in section 4.1, for example the application to predictive influence functions, the methods of section 3 do not apply directly to the entities of interest (the PIF's), but do supply useful information about the approximation. As for the efficiency with which the available information for the model and data at hand is used, the integrals method in its current implementation runs very quickly. The contours method is much slower, because the current version numerically differentiates a function that is evaluated by solving an equation numerically. In the next version of the program that implements the methods, this problem will be circumvented by using a formula for dr_h/dh, derived by implicit differentiation, that only requires differentiating the exact function g, and then the contours method should run as quickly as the integrals method.

Both methods, as defined in section 3, are calibrated as probabilities, which are numbers about which statisticians at least can be presumed to have cultivated some intuition. The contours method extended to likelihood regions (section 4.2) is calibrated in terms of nominal confidence coefficients, about which the same judgement should apply.

Finally, the methods of section 3 do not rely on approximations at all for

their definitions, although approximations are necessary for their computation.

While the methods presented here seem to satisfy the requirements of section 2.3 better than convergence rates or the geometric approximation of section 2.2, their ultimate utility will depend on two things: the ease with which they can be computed relative to the ease of computing the exact quantities or functions of interest, and their theoretical and intuitional fertility. It remains to be seen whether the two new methods will satisfy these last requirements.

Acknowledgements

References

Bartlett, M. S. (1953). Approximate Confidence Intervals. Biometrika, Vol. 40, pp. 12-19.

Bartlett, M. S. (1953). Approximate Confidence Intervals. II More Than One Unknown Parameter. Biometrika, Vol. 40, pp. 306-17.

Bates, D. M., and D. G. Watts (1980). Relative Curvature Measures of Non-linearity (with discussion). JRSS Ser. B, Vol 42, No. 1, pp. 1-25.

Bates, D. M., and D. G. Watts (1981). Parameter Transformations for Improved Approximate Confidence Regions in Non-linear Least Squares. Ann. Statist., Vol. 9, No. 6, 1152-67.

Beale, E. M. L. (1960). Confidence Regions in Nonlinear Estimation (with discussion). JRSS Ser. B, Vol. 22, pp.41-88.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press.

Cook, R. D., and J. A. Witmer (1985). A Note on Parameter-Effects Curvature. JASA, forthcoming.

Cox, D. R., and D. V. Hinkley (1974). Theoretical Statistics. London: Chapman & Hall.

Durbin, J. (1980). Approximations for Densities of Sufficient Estimators. Biometrika, Vol. 67, No. 2, pp. 311-33.

Efron, B. (1975). Defining the Curvature of a Statistical Problem. Ann. Statist., Vol. 3, 1189-1242.

Fisher, R. A. (1973). Statistical Methods and Scientific Inference, 3rd ed. New York: Hafner.

Gallant, A. R. (1975). Testing a Subset of the Parameters of a Nonlinear Regression Model. JASA, Vol. 70, pp. 927-32.

Geisser, S. (1970). Bayesian Analysis of Growth Curves. Sankhya, Series A, Vol. 32, Part I, pp. 53-64.

Hamilton, D. C., D. G. Watts, and D. M. Bates (1982). Accounting for Intrinsic Nonlinearity in Nonlinear Regression Parameter Inference Regions. Ann. Statist., Vol. 10, No. 2, pp. 386-93.

Hodges, J. S. (1985). Methods for Assessing the Accuracy of Some Standard Approximations. Ph.D. dissertation, University of Minnesota.

Jennings, D. E. (1982). Inference and Diagnostics for Logistic Regression. Ph.D. dissertation, University of Minnesota.

Johnson, N. L., and S. Kotz (1970). Continuous Univariate Distributions. New York: Houghton-Mifflin.

Johnson, W., and S. Geisser (1982). Assessing the Predictive Influence of Observations. In G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh, eds., Statistics and Probability: Essays in Honor of C. R. Rao. New York: North-Holland.

Johnson, W., and S. Geisser (1983). A Predictive View of the Detection and Characterization of Influential Observations in Regression Analysis. JASA, Vol. 78, pp. 137-44.

Lee, J. C., and S. Geisser (1972). Growth Curve Prediction. Sankhya, Series A, Vol. 34, Part 4, pp. 393-412.

Lee, J. C., and S. Geisser (1975). Applications of Growth Curve Prediction. Sankhya, Series A, Vol. 37, Part 2, pp. 239-56.

Lindgren, B. W. (1976). Statistical Theory. New York: Macmillan.

Luenberger, D. G. (1984). Linear and Nonlinear Programming, Second Edition. Reading, Mass.: Addison-Wesley.

McCullagh, P., and J. A. Nelder (1983). Generalized Linear Models. London: Chapman & Hall.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). Multivariate Analysis. New York: Academic Press.

Ratkowsky, D. A. (1983). Nonlinear Regression Modelling. New York: Marcel Dekker.

Weiss, R. (1985). Personal communication.

Welch, B. L., and H. W. Peers (1963). On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods. JRSS Ser. B, Vol. 25, pp. 318-29.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. New York: John Wiley & Sons.